Performance Metrics for Intelligent Systems
An Engineering Perspective

Rong Gao and Lefteri H. Tsoukalas
Purdue University
School of Nuclear Engineering
Applied Intelligent Systems Laboratory
W. Lafayette, IN 47907-1290
gao@ecn.purdue.edu, tsoukala@ecn.purdue.edu

ABSTRACT

This paper provides a general discussion on performance metrics for intelligent systems drawing largely on our experience with engineering applications. The experience has led us to view machine intelligence as a type of machine-facilitated human intelligence. This view implies that the locus of machine intelligence is to be found in relations amongst humans *vis a vis* the machine (in the subject-to-subject-via-object relation). Hence, quantitative metrics for intelligence may be sought as functions of the human-machine and machine-machine interface; evaluating them may be achieved through a conventional behaviorist-type of approach where a system is characterized by observing its response to given inputs. Some guidelines for this process that we found to be useful are discussed in the paper.

Keywords: *Intelligence, Performance Metrics, Fuzzy Logic, Neural Networks*

1. INTRODUCTION

Since the onset of industrial revolution machines have been developed to relieve humans from tedious work and reduce the cost of producing goods and services. With the advent of the digital computer, the industrial notion of machines as artifacts capable of mechanical work has been greatly expanded to include sophisticated capabilities of information processing, decision-making, communications, sensing, coordination and control. Somewhat naively but, as we will argue latter on, quite accurately, we tend to refer to this post-industrial sophistication as "machine intelligence."

Equipped with control and computing units, machines can be made to be very predictable and quite reliable. The reliability is exemplified by the fact that machines follow given instructions literally regardless of changes in their environment. This characteristic is undesirable in complex applications. Predefined instructions are usually coded by human experts and make it nearly impossible to take into account every possible scenario that might happen in real-world applications. Therefore, it is desirable that machines possess capabilities for appropriately handling cases that they have not been directly instructed. Although there have been numerous definitions of intelligence, engineers tend to think that a machine or a system in general is intelligent if it is capable of handling exceptions properly. In this respect the engineering view is somewhat similar to that of the cognitive scientists who look at various *agnosias* (for example, *visual agnosia* or failure to recognize objects seen) as opportunities and tools for understanding the complex inner-workings of the human brain (Gazzaniga, 1998). But, the engineering aim is quite different. Intelligent systems are transforming the way we design, fabricate, operate and even dispose complex engineering artifacts such as airplanes and power plants. A Boeing 777 airliner, for example, may have in excess of nine million distinct parts, while the number for an advanced boiling water reactor is the 109 range. Intelligent systems are necessary to make such systems safe, economical and manageable at all times.

It should be noted that the adjective "intelligent" gradually fades away from designating any system that becomes routinely available and widely familiar. Intelligence is an attribution reserved for systems that are at a more nascent level of development and more likely not proven or established technology. Building intelligent systems is a goal that often appears to be quite elusive. Hence, having a metric for intelligence, any metric, is useful not only for comparing system A to system B, but also, for comparing system A at an early age of development to system A at more mature level of development. In this respect an index for intelligence is not different from any metric of performance that can be consistently applied to assess the growth of a system. This a very important issue for systems such as nuclear power plants or passenger airliners whose lifespan may be comparable or exceeding the lifespan of their designers and operators.

We have observed in numerous engineering applications that a major difficulty giving rise to the elusiveness of machine intelligence is due to our deeply held notions and assumptions about machines. Humans seem to be so overwhelmingly prepared to think of machines as quite independent entities, separate from us; distinct and also distinguished; sitting outside the boundaries of human boundaries (physiologically, cognitively, and socially); and yet so intimately ours. Machines are always adjuncts to humans. Although we describe them in terms of objective qualities (such as power, mass, volume) their most important attributes are the ones relating to their functionality and purpose (interfacial characteristics). Intelligence is a functionality not an objective property. Yet, and because of that, machine intelligence involves what the psychologists call *reification*, that is, something appears to exist just because we have word(s) for it.

The view we espouse is that machines, including the sophisticated and computationally savvy artifacts of today and tomorrow, are accessories to human intelligence. They have no intelligence of their own (to have that they would have to live the lives of humans). Intelligent systems are machines functioning as a medium for playing out the drama of human intelligence; principally activities for asking and answering questions, a kind of generalized dialogue amongst humans (not anymore constrained to be physically present). Our view defines intelligent systems as *virtual interlocutors*, that is as systems that function in a way that makes it possible for humans and other machines to ask and answer questions unconstrained by personal presence or awareness. In this sense, machines can be viewed as "intelligent" although we all know that they could not possibly come about on their own, without human volition, know-how, design and material implementation.

Viewing intelligent machines as "virtual interlocutors," raises the question of language. What is the right idiom for the man-machine discourse we are talking about? It has to be a language that its ultimate aim is to facilitate a virtual dialogue amongst humans and as such it is desirable to have the computational characteristics of natural language. For this, we have to turn to fuzzy logic. It is extremely difficult to capture within any formal language the complex and rich attributes of natural language including, but not limited to, flexibility, semantic depth, computational economy (parsimony), and portability. We strongly believe that fuzzy logic is a highly promising tool; its potential is largely uncapped, its full power is still to be harnessed.

Additional frustrations with intelligent machines are caused by the lack of a bridge that links any interpretation of intelligence (such as the one put forward in this paper) to the implementation of intelligence. Definitions of intelligence do not often provide useful information needed by engineers in the realization process. Engineers would like to have quantitative performance metrics that could be used to measure the degree of intelligence of a system. Defining a performance metric is, however, not easier than building an intelligent system itself. Theoretically, provided a quantitative performance metric is available, methods can be developed to optimize the system's performance to reach a threshold that this particular system is deemed intelligent. We have gathered plenty of experience in designing optimization algorithms and they are being used in a variety of applications involving neural networks and fuzzy logic (Tsoukalas, 1997). The research on performance metrics for intelligent systems is an important focus. In the following sections we present a general discussion and some guidelines for designing performance metrics.

2. INTELLIGENT SYSTEMS

Researchers of artificial intelligence have traditionally defined intelligence as an inherent property of a machine. An intelligent machine (or system) is viewed as one that has some computational capacity to act like a human, that is, "think" humanly, or "act" rationally, or "think" rationally (Russel, 1985). Hence, computational metrics of intelligence are traditionally expected to measure how well a machine performs like a human, for example, like a chess master, or like an expert diagnostician.

We believe that such thinking-like-human approaches trying to mimic the way a mind operates are not technically feasible with current engineering capabilities. First, the complexities of human brain make simulation impossible using existing technology. The latest Intel Pentium IV processor integrates 55 million transistors, much less than the 100 billion neurons and 100 trillion synaptic junctions found in a person's brain. Second, the basic processing unit in a computer system, the transistor, is identical throughout a processor and can only handle two numbers, 0 and 1. On the other hand, neurons are diversified and are capable of processing subtle electrochemical signals efficiently in numerous possible ways. Third, humans as complex biological systems are the result of millions of years of natural selection. Many species coexist but only humans have emerged as intelligent creatures (in a full sense of the word). And certainly part of their story is found in language and their capacity to form complex cultural and technical artifacts and social institutions.

We think of intelligence as an advanced functionality of a system. Based on the discussion above it is quite clear that this type of functionality should be independent of a machine's internal implementation details. We propose that this functionality is a function of the interface (human-machine, machine-machine); to be found in the subject-to-subject relation *vis a vis* the object, that is, the computer or "intelligent machine." In order for the intelligent functionality to be something observable (measurable), the intelligence of a system ought to be judged based on a system's response to provided inputs. The internal structure and implementation may be not so important for the purpose of observing it.

But, is it really possible to measure intelligence quantitatively? We believe that the answer ought to be yes; else, we run the risk of viewing machine intelligence as something metaphysical, mysterious and therefore not amenable to investigation. Methodologically, we know that it is very hard to quantify machine intelligence. The reason is that any intelligence measure has to be an overall performance index involving many detailed measures. For example, an intelligent person may be extraordinarily good at math but very poor in music. For another person the opposite may be true (capable in music but incapable at math). Obviously there is no single number that can be used to characterize the difference between these two persons. Human IQ tests have been criticized for their non-typicality, unreliability and inconsistency. The incommensurability of intelligence is a barrier we have to face with intelligent machines as well; and not only because they are used or they are better at different things. Evidently, it is different humans that make for very different intelligent machines.

Despite all these difficulties, defining performance metrics for intelligent systems is a worthy goal. People realize that the long lack of universally acceptable measures has seriously hampered the process of intelligent systems development. Science and technology have advanced by cooperation and competition. The root of cooperation and competition is a common ground with which results of different researchers can be compared. Comparisons are impossible without agreement on performance metrics.

Some philosophical and methodological barriers to intelligent metrics can be overcome by adopting a pragmatic approach. Such an approach focuses on the specifics of the problem and calls for strategies for improvement (learning) and development (maturation) within a given context and with well-defined metrics. Thus, for the first step a pragmatic approach is to identify the needs of the applications. What are the situations that intelligent systems are designed for? Are we going to develop an intelligent system that clones a human being (in some way) or solve a specific problem (for the benefit of well-defined user needs)? Undoubtedly, the efforts involved in designing performance metrics for these two different systems are rather incomparable.

3. General Guidelines

In this section we present five items we found important in designing performance metrics. They are intended to be general guidelines. A comprehensive performance metric should take into account all five of the proposed items. The first describes a questionnaire-based method similar to approaches taken for knowledge solicitation in well-defined application domains. The second identifies the capacity for generalization. The third item identifies the need for adaptation. The fourth captures the social or group capabilities of intelligent systems. Finally, the fourth identifies transitivity or how different intelligent systems ought to be comparable.

3.1 General and Specific Metrics

Detailed quantitative metrics of general intelligence are difficult to formulate and potentially not necessary. Intelligence in general integrates so many parameters and is not possible to have an objective general measure. However, approximate and application oriented measures are possible. It is a lot easier to develop a metric to evaluate the performance of some system for a specific application like chess playing or medical diagnosis. Therefore, if possible, application-specific measures should be always considered first. Application-specific measures can be constructed based on a **set of questionnaires.** Techniques from knowledge solicitation and web-based assessments can be used. Questionnaires can be analyzed via statistical approaches or fuzzy quantification (Tsoukalas, 1997).

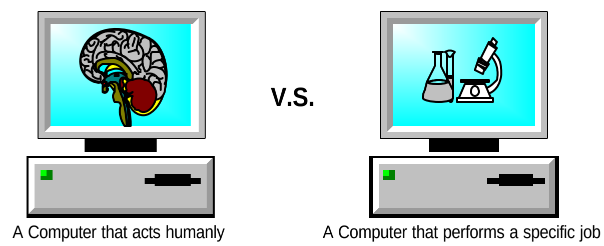

Figure 1. General intelligence and specific intelligence

3.2 Metric for Generalization Capabilities

The degree of intelligence ought to reflect in some fashion the capacity for generalization. An intelligent system solves a problem first by searching its previous experience for similar cases. The first level of intelligence is looking for a direct match. A higher level of intelligence is needed when a direct match is unavailable. The higher level of intelligence appears as the capability of maneuvering experience in part in order to generate new unseen instances, which resemble the problem to solve, as shown in figure 2. Suppose, for example, that the problem is to classify some unknown shape. The direct match is the first approach and essentially compares the given shape against ideal geometrical shapes. The indirect match may do the matching against generalized instances and more generally against composite generalized instances. An appropriate metric in this example ought to capture the ability of the system to deal with the more general topological transformations involved in the indirect matches. Although a lot has been written for generalization, typically generalization in many engineering applications is little

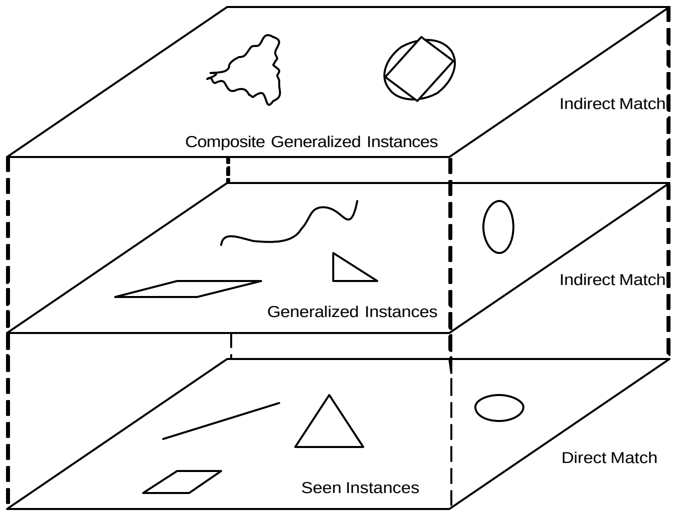

Figure 2. An illustration of different levels of generalizations

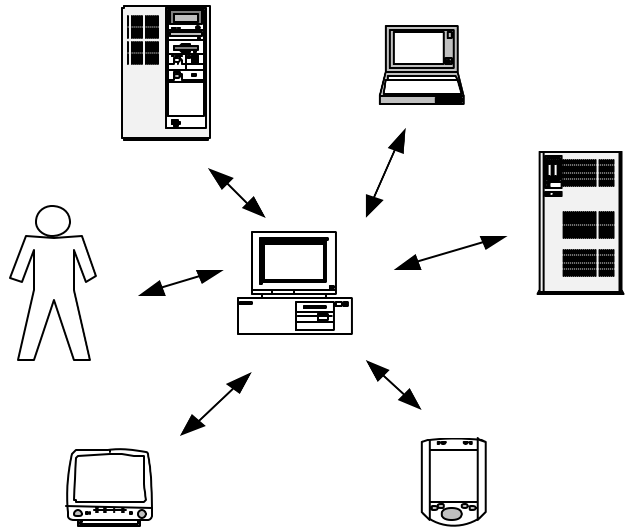

Figure 3: System and its environment

different from interpolation. But, that's fine. Even a good interpolation metric is adequate and very useful if applied consistently as a measure of generalization.

3.3 Intelligence Metric should be Adaptive

The intelligence of a system cannot really be evaluated by a fixed rule. Rather it ought to be a collective index that reflects the overall performance of the evaluated system on a variety of situations. Consequently, this metric should be dynamic and adaptive. It should change to adapt to the new information that has been gathered regarding a system's more recent performance. In this respect a simple neural network can be very useful as means of adaptation of the metric (even though the range of adaptation may be rather narrow).

3.4 Intelligence as a Social Characteristic

From an engineering viewpoint, an isolated system, no matter how intelligent it is, is not of great interest. Measuring intelligence should be performed in the context of a group or society that includes other systems (computers or humans), as shown in figure 3. Any intelligent ability ought to be evaluated from the interrelations among multiple systems. In a sense the kind of machine intelligence we are called to quantify almost always involves network systems, be they computers, sensors, robotic devices, controllers, expert systems, or search engines in the Internet. The criteria for judging the social abilities of interacting systems are the correctness of interpretations of their inputs and the effectiveness of presentation of the outputs.

If a system is put into a society, we have to consider its relations with other parties. In a society, a system always stores in its local database the profiles of other systems that it knows. In order words, the intelligence profile of a system is distributed to the society. The argument that an intelligence metric should be dynamic and adaptive requires that a system be constantly monitored. Assigning a dedicated agent to perform this job will be biased and unreliable. The solution is that any system ought to be examined by its peers (the rest of the systems in the same society). This implies that an intelligent system (human or computers) should posses the ability to evaluate the intelligence of other systems. However, it is impossible for one system to evaluate all other systems directly, partly because of security reasons or simply because of the exponentially growing communication overhead. In such cases, one system needs to reach its own judgment indirectly based on the judgment (which may be direct or indirect one) of other systems. For example, in figure 4, System 1 is about to evaluate the performance of System 3. However, System 1 has no means to communicate with System 3 directly. In such a scenario, it is possible for System 1 to reach a decision based on the evaluations obtained from System 2 and System 4 (with which it has direct connections). The third party information might be direct (such as System 2 that has direct connection with System 3) or indirect (such as System 4, which in return relies on System 5 to make its own decision). To achieve this type of feature, a system needs to "trust" to some degree the capabilities of other systems.

3.5 Transitivity

Finally any intelligence metric should be used with care, especially when comparisons are involved. The direct comparison between two systems using an intelligence metric

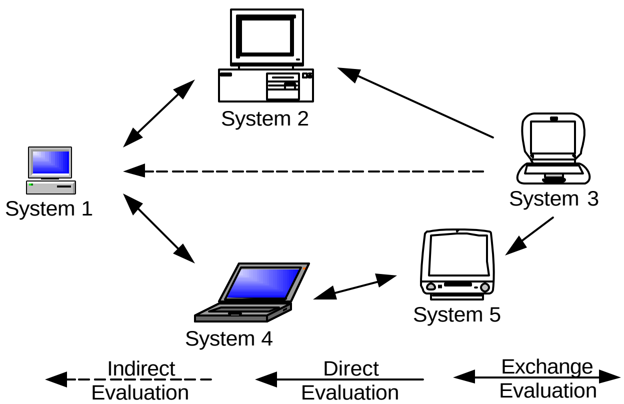

Figure 4. Indirect evaluation of intelligence

remains questionable, except when this metric has been defined in a very narrow sense such as for the performance of a specific job such as medical diagnosis. Generally, we should avoid the use of an intelligence metric in a chained fashion because of its inherent uncertainty and incommensurability. For instance, system A is more intelligent than system A, which is more intelligent than system C. Yet, it is not correct to imply that system A is more intelligent than C.

4. A PROTOTYPE FOR GENERAL MEASURES

An intelligence metric in general is difficult to be written in an analytical form. However, an engineering construction approach may be useful. The process of evaluating intelligence itself is an intelligent process. To break this infinite loop, we must start from some systems that are canonically intelligent. Humans are the main option. Some prototype machine systems are first constructed and are approved to be intelligent by humans. These initial systems are not necessarily perfect in terms of natural intelligence. The criteria of intelligence are numerous and the most important one is the capacity of judging the intelligence of other systems. The intelligence evaluation process is not a calibration process where a less precise machine is able to calibrate a more precise one. A more pragmatic approach is needed. The prototypes that have been evaluated by humans now can evaluate other machines, as shown in figure 5. The evaluated systems can be further used to evaluate other systems. It should be noted that this is not a one shot operation and the first round of evaluation is usually inaccurate. In later iterations, every system (including the first prototype) will have the chance to improve its evaluation capability by comparing others' evaluations with its own one. A steady state for the system, if reachable, informs us that it has achieved stable and more accurate evaluations for other systems. The key of this approach is that an intelligent system is able to talk to its neighbors.

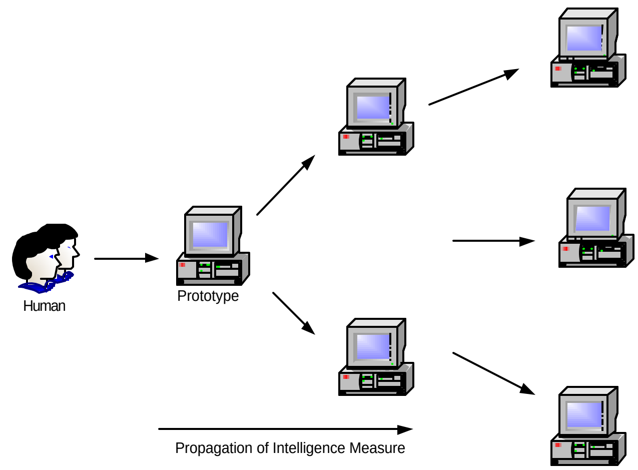

Figure 5. Construct intelligence metric in a networking approach

5. CONCLUSIONS AND REMARKS

We have discussed several important guidelines for intelligent metrics. The key is to focus on the interface (machine-machine, machine-human). A system's intelligence is reflected by the ways of processing inputs and presenting outputs. The interface not only accepts problem-solving data but only those "control" or "judgment" pieces of information, such as evaluations by other systems.

Over the years there has been a great interest in constructing intelligent systems not only because machines can potentially solve problem more consistently and flexibly (with human supervision) but also because intelligent systems are a great metaphor for intelligent human activities. The activities of posing and answering questions and of building knowledge through a dialectic process are now greatly facilitated by computer systems which we tend to view as "intelligent." The results are an unprecedented and much needed access to the human mind. And, machines that not only surprise and fascinate us but, most importantly, we cannot do without them; in the sense that we cannot really manage the complexity of exceeding complicated engineering artifacts that come to be over several generations.

However, there is no free lunch. The associated cost is the so called *responsibility dilemma*. A system is intelligent because it is capable of handling exceptions that it was never taught. The question then may be raised as to who should be responsible for the consequence incurred by the "intelligent actions." The designer of the system should not be blamed because the system does not follow instructions literally and the designer can not foresee the direction that the system evolves. These issues have to be addressed and if possible reflected in future intelligence metris.

6. REFERENCES

Gazzaniga, M.S., Ivry, R.B., Mangun, G.R., *Cognitive Neuroscience: The Biology of the Mind*, W.W. Norton & Company, New York, 1998.

Tsoukalas, L. H, and Uhrig, R. E., *Fuzzy and Neural Approaches in Engineering*, John Wiley and Sons, NY, 1997.

Russell, S. and Norvig, P., *Artificial Intelligence: A Modern Approach*, Prentice Hall, New Jersey, 1995

7. BIOGRAPHY

Rong Gao obtained his bachelor's and master's degrees from Tsinghua University, Beijing in 1994 and 1997 respectively. In December 2000, he received his master degree in electrical and computer engineering, Ph.D. in nuclear engineering, both from Purdue University. He is currently working at the Artificial Intelligence Systems Laboratory (AISL) as a visiting Assistant Professor. His research interests include human-machine interface, intelligent signal processing and complex systems modeling.

Lefteri H. Tsoukalas is Professor and Head of the School of Nuclear Engineering at Purdue University and the Director of the Applied Intelligent Systems Laboratory (AISL). He has a PhD from the University of Illinois (1989) and numerous publications in neurofuzzy systems including the book *Fuzzy and Neural Approaches in Engineering*, co-authored with R.E. Uhrig (Wiley, 1997).